SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction

Abstract

Recently video generation has achieved substantial progress with realistic results. Nevertheless, existing AI-generated videos are usually very short clips ("shot-level") depicting a single scene. To deliver a coherent long video ("story-level"), it is desirable to have creative transition and prediction effects across different clips. This paper presents a short-to-long video diffusion model, SEINE, that focuses on generative transition and prediction. The goal is to generate high-quality long videos with smooth and creative transitions between scenes and varying lengths of shot-level videos. Specifically, we propose a random-mask video diffusion model to automatically generate transitions based on textual descriptions. By providing the images of different scenes as inputs, combined with text-based control, our model generates transition videos that ensure coherence and visual quality. Furthermore, the model can be readily extended to various tasks such as image-to-video animation and autoregressive video prediction. To conduct a comprehensive evaluation of this new generative task, we propose three assessing criteria for smooth and creative transition: temporal consistency, semantic similarity, and video-text semantic alignment. Extensive experiments validate the effectiveness of our approach over existing methods for generative transition and prediction, enabling the creation of story-level long videos. Project page: https://vchitect.github.io/SEINE-project/ .